SCHOOL OF SUSTAINABLE ENGINEERING AND THE BUILT ENVIRONMENT

No Right Answer:
A Feasibility Study of Essay Assessment with LDA

Tom Roberts

ASU-SSEBE-CESEM-2013-RPR-005
Research Project Report Series

April 2013

No Right Answer: A Feasibility Study of Essay Assessment with LDA

Tom Roberts
April 2013

Abstract

Essay scoring is a difficult and contentious business. The problem is exacerbated when there are no "right" answers for the essay prompts. This research developed a simple toolset for essay analysis by integrating a freely available Latent Dirichlet Allocation (LDA) implementation into a homegrown assessment assistant. The complexity of the essay assessment problem is demonstrated and illustrated with a representative collection of open-ended essays. This research also explores the use of "expert vectors" or "keyword essays" for maximizing the utility of LDA with small corpora. While, by itself, LDA appears insufficient for adequately scoring essays, it is quite capable of classifying responses to open-ended essay prompts and providing insight into the responses. This research also reports some trends that might be useful in scoring essays once more data is available. Some observations are made about these insights and a discussion of the use of LDA in qualitative assessment results in proposals that may assist other researchers in developing more complete essay assessment software.

Introduction

Experimenting with new ideas is an important feature of learning. Despite inconsistent and conflicting empirical support, it is still thought by many that among the best ways to grapple seriously with concepts and learn to effectively communicate is through writing about them (Klein, 1999; Kieft et al., 2006). In fact, in a provocative and clever move to make this argument, Pearson (*pearsonkt.com*) has named its essay tutoring package *WriteToLearn*. Universities frequently offer classes in which students write essays to demonstrate their grasp of concepts and their ability to explain, critique, integrate, augment, and extend important ideas. Generally, the scores of these essays contribute to a student's overall assessment for the course. When such classes are large, essay scoring can be a daunting task for the teaching staff and automated essay scoring (AES) dangles a tempting carrot in front of any such team. But AES has received mixed reviews in the literature, with strong proponents (usually selling products) facing off against equally strong opponents arguing for the purity of a craft that cannot be assessed by a machine (Deane, 2013; Perelman, 2012).

The problem is exacerbated when the required essays address open-ended prompts that have no single correct answer. While some might suggest that an essay approach can never legitimately expect a single "right" answer, it is easy to see the difference between asking a student to discuss the causes of the US Civil War (which *must* include a discussion of the economics of slavery to be "right") and one that asks a student to predict the ascendant culture in 100 years. In the latter case, there are certain elements and trends that must be addressed in a high quality essay, but depending on the manner in which the essay is approached, there can be many wildly divergent and equally defensible arguments.

This research was prompted by such a situation and the resulting scoring dilemma faced by the instructional team. In a series of courses (ASU's CEE 181, 400, 581) designed to demonstrate the complexity of the world of the future and equip engineers to face it, the assignments require students to interact with complex ideas through essay writing. But the dilemma of the teaching team extends far beyond simply assessing the quality of the essays. The forces to resolve in the grading problem can be grouped into four major categories as shown in Figure 1.

First, for such expansive and open-ended topics, teaching assistants (TAs) tend to need multi-disciplinary training and more life experiences in order to perform well in both classroom discussion and essay grading duties. Further, with the growing class size, it is difficult to find and train enough skilled TAs willing to invest the required time. Second, the course themes are provocative and require integration of ideas across multiple disciplines—a skill engineers need now more than ever. In discussing the impact of technology on culture (with all its tentacles, including political discourse, religion, economics, etc.), the courses expose a category of problems referred to by Rittel & Webber (1973) as "wicked." This leads to, third, a group of essay prompts that have decidedly no correct answer, but still require intelligently structured responses. Finally, for the assignments to be effective, students must receive high quality feedback that not only points out their communication issues (grammar, spelling, thesis, sentence and paragraph structure, argument flow, etc.), but also steers their growth in acquiring the content they need to address such complex problems.

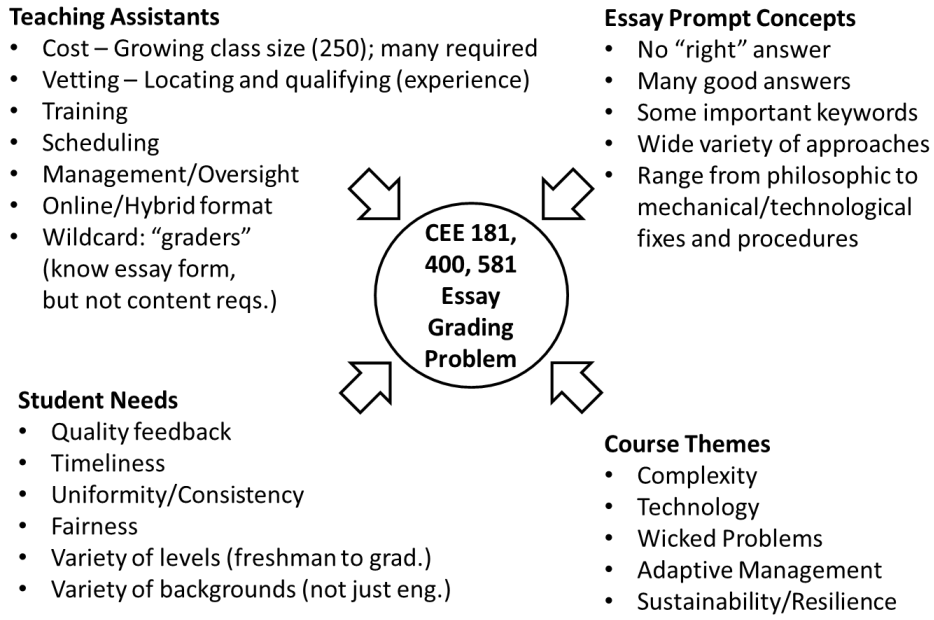

Figure 1. Forces to resolve in the CEE 181, 400, 581 Essay Grading Problem

Given this complicated array of forces, the primary research question centered on whether or not a software package could in any way alleviate the burden of the teaching team in scoring the essays for this suite of courses. While a one-size-fits-all package might still be a pipedream, and machine scoring still has significant hurdles to overcome, it is clear that software tools *can* assist in the effort.

As might be expected, there were important qualifications on that goal including the need to contain costs and limit impact to the team. For example LightSIDE labs (*lightsidelabs.com*) of Pittsburgh, PA, a 2013 startup company specializing in essay assessment, boasts "instant feedback to students" and "lightened workloads for graders" but the service they provide requires a minimum startup fee of $2,000 and a regular monthly service charge of $500 (*lightsidelabs.com/enterprise*) for essay grading. Even given what could be considered a very reasonable cost structure, adopting LightSIDE's solution requires a commitment and a long term investment. Indeed, most of the solution they offer is available for download free of charge, but requires significant software expertise and system management support to make it operational.

The Waikato Environment for Knowledge Analysis (Weka) is a popular workbench of machine learning software developed at the University of Waikato, New Zealand. Weka is also free software available under the GNU General Public License (see *fsf.org* and *gnu.org/licenses/gpl.html*), but comes with a significantly high price tag when integration time and learning curve are considered. Toolsets like Weka (upon which many tools—including the LightSIDE solution—are layered) are costly in both time and resources. They are time consuming for system integration, learning curve, training, and tuning. They are resource intensive because they require special equipment (servers and workstations), and trained human staff in the form of webmasters, and IT administrative support. Further, they are workflow-altering—an impact that is sometimes rewarding but only if the package is fully adopted and becomes a "way of life" for the instructional team. All of these are invasive and costly and therefore not suitable for instructional teams that are attempting to do more with less (for the good of the student!) during a period of budget cutbacks.

Hence, an important goal of the research was to explore what software elements might be freely available for integration into a minimalist package, specifically optimized for the needs of the instructional team. As a consequence the specific contributions of this research include:

- Demonstration of integration, exploitation, and improvement of open source software,
- Development of a dashboard interface for the instructional team that fits the current workflow,
- Exploration into optimizing the value of small training sets,
- Evaluation of potential use of "expert vectors" (short, bulleted, keyword-only "essays") as training tools, and
- Demonstration of the utility of Latent Dirichlet Allocation (LDA) in both qualitative and quantitative assessment of essays.

Specifically, if a software package could be developed with a part-time focus in a two-month period that provided assistance of any kind, it was worth attempting. That assistance could even be in the form of better understanding the data set and learning how much new data would be needed in the future. Such was the goal of this research effort.

Background

A hugely contentious issue (Ben-Simon & Bennett, 2007; Condon, 2013; Grimes & Warschauer, 2010; Patterson, 2005; Vojak et al., 2011), Automated Essay Scoring (AES) is starting to see widespread use in high-stakes grading, ranking, and qualification exams (GRE, GMAT, TOEFL, etc.). Because of this, the controversy has reached beyond academia into the public square (Mathews, 2004). The debate underlying application of AES belies a much deeper argument over what is actually being assessed. As Perelman (2012) and Deane (2013) argue, it really boils down to the "definition of the writing construct" (Deane, 2013, p. 9) and disagreement over whether something so entirely "human" can ever be properly assessed by a machine.

It is helpful to think of essay scoring systems in *tiers* or layers of progressively complex function (cf. ETS, Pearson & The College Board, 2010):

- Tier 1, where most current automated scoring packages reside, targets aspects of grammar, usage, mechanics, spelling, and vocabulary. Products at tier 1 are generally well-positioned to score essays which are intended to measure text production skills. Tier 1 packages are frequently helpful in such courses as the archetypical "English 101" which trains, practices, and measures such skills.

- Tier 2 systems add evaluation of the semantic content of essays, relevance to the prompt or suggested topic, and aspects of organization and flow. Far fewer tools are in this category, two of which are discussed below.
- Tier 3 systems (of which none exist) will (eventually) assess creativity, figures of speech, poetry, irony, or other more artistic uses of writing. These systems might also assess rhetorical voice, logic of an argument, extent to which particular concepts are accurately described, or whether specific ideas presented in the essay are well founded. Depending on how the "construct" of writing is defined (Deane, 2013), these are arguably the most difficult and most important areas for essay scoring—and this is the basis of the argument against a machine's ability to ever adequately assess the written form.

ETS's *Criterion* is an essay-writing tutor at Tier 1+ that appears to focus more on form than it does on content. Criterion's *e-rater* provides automated scoring of writing quality including:

- errors in grammar (e.g., subject-verb agreement),
- usage (e.g., preposition selection),
- mechanics (e.g., capitalization),
- style (e.g., repetitious word use),
- discourse structure (e.g., presence of a thesis statement, main points), and
- vocabulary usage (e.g., relative sophistication of vocabulary) (ETS, 2013a).

ETS's *c-rater* is purported to provide automated scoring of written *content*, but note that this differs from the content assessments of Tier 2 systems outlined above. ETS defines their content assessment as including:

- correcting student's spelling,
- determining the grammatical structure of each sentence,
- resolving pronoun reference, and
- reasoning about words and their senses (ETS, 2013b).

Despite calling them "content", these items seem to reside squarely in the "form" camp. Note that ETS specifically differentiates its so-called "deep linguistic analysis" from "purely statistical approaches based on words, such as latent semantic analysis (LSA)" indicating that including grammatical information as they do (and by implication, as their competitors do *not*—since LSA ignores grammar) reduces the chance that students will be misled by the assessments (ETS, 2013b) if, say, their vocabulary was solid and relevant while their grammar was deficient. While ETS remains reticent about sharing their technological approach, it appears they are exploring something beyond simple unigram semantic models.

Criterion can be compared to Pearson's *WriteToLearn* product which embeds the *Intelligent Essay Assessor* (Landauer, Laham & Foltz, 2003). This product originally focused on Tier 2 functions (content) but integrates other tools to deliver its tier 1 functions. To accomplish the content analysis IEA employs Latent Semantic Analysis (LSA), "a machine-learning model of human understanding of text" (Landauer et al., 2003, p. 297). They continue:

> While the LSA model of verbal meaning at first appears to be an implausible over-simplification, it turns out to yield remarkably accurate simulations of a wide spectrum of language phenomena, and robust support for automation of many language-dependent tasks (p. 297).

Obviously, in the ten years since the cited publication, Landauer et al. have significantly augmented their self-styled "implausible" yet "remarkably accurate" approach. This likely includes expansion into *n*-gram semantic models, but they are equally reticent about sharing their approach.

While AES approaches are many and varied, it is clear that based on these two arguably state-of-the-art, commercially available systems, the writing construct's easy decomposition into "form" and "content" has forced their convergence from disparate paths. That is, ETS's *Criterion* may have started with a focus on form, but has recently added content to the mix, while Pearson's *WriteToLearn* started in the content analysis world (with IEA) and has since added form to its marketable bag of tricks.

This distinction between form and content is important to the CEE 181/400/581 instructional team. Our goal is not specifically to provide what English 101 should have given to students (form skills). Though we want all our students to be good writers and effective communicators, it is more important to us that they can integrate complex ideas and concepts (content skills). Though we want clear, concise and eloquent prose that makes a point and argues persuasively, we are more interested in credible *content* than we are in the *form* of the essay. Still, since these *are* college students, we demand that the form be at least adequate. We rely on English 101 to deliver students who are adept at text production, and we hope to take them to next level. Not only do we want them "to be able to say it," but we want them "to have something important to say." This need for balance—and the huge grading burden—has driven the team to seek something that can assist in measuring the *content* of the essays.

Exploring Latent Dirichlet Allocation (LDA)

This research chose to explore the potential value of Latent Dirichlet Allocation (LDA) in assessing essay content for the CEE 181/400/581 suite of classes. LDA is a bag-of-words approach that is similar to the LSA approach employed by IEA, but it has proven to be demonstrably better at topic elicitation and classification tasks (Chang et al., 2009; Blei, Ng & Jordan, 2003). While LDA has important applicability in a broad range of data processing functions (including image processing), it is used in this effort strictly in processing text: words, documents, corpus (a collection of documents), and corpora (multiple collections).

In a nutshell, documents contain a mixture of topics, and topics are represented by groups of words. What I refer to herein as an "LDA Estimate" or a "model" is a generative probabilistic model of a collection of documents (a corpus)—the probabilities that certain words in the documents are associated with certain topics discussed in those documents. More formally, documents (essays) are probabilistic collections of latent topics (with measures of degree), and a topic is a distribution of words. As Blei et al. (2003) put it, "LDA posits that each word of both the observed and unseen [(held out)] documents is generated by a randomly chosen topic which is drawn from a distribution with a randomly chosen parameter" (Blei, et al., 2003, p. 1002).

Figure 2. LDA Overview provides an overview of the mechanics of the LDA concept. It shows a collection of essays which represents a typical corpus for which an LDA model is to be generated. Prior to generating the model, some documents are held out (solid line) so they can later be used to "inference" against the model. The LDA software then processes the corpus and generates a model with a user-specified number of topics (illustrated here with six, numbered 0 through 5). This is the estimation process. Once the model is generated (estimated), the held out essay(s) can be processed (dashed lines) and the LDA software will assign probabilities of fit to each inferred topic (illustrated here with 12%, 33%, 21%, etc.). The highest probability allocation (e.g., 33%, shown in a dashed circle) generally indicates the best match of topic keywords. It is these allocation probabilities (think of it as a degree of match) that are used throughout this analysis.

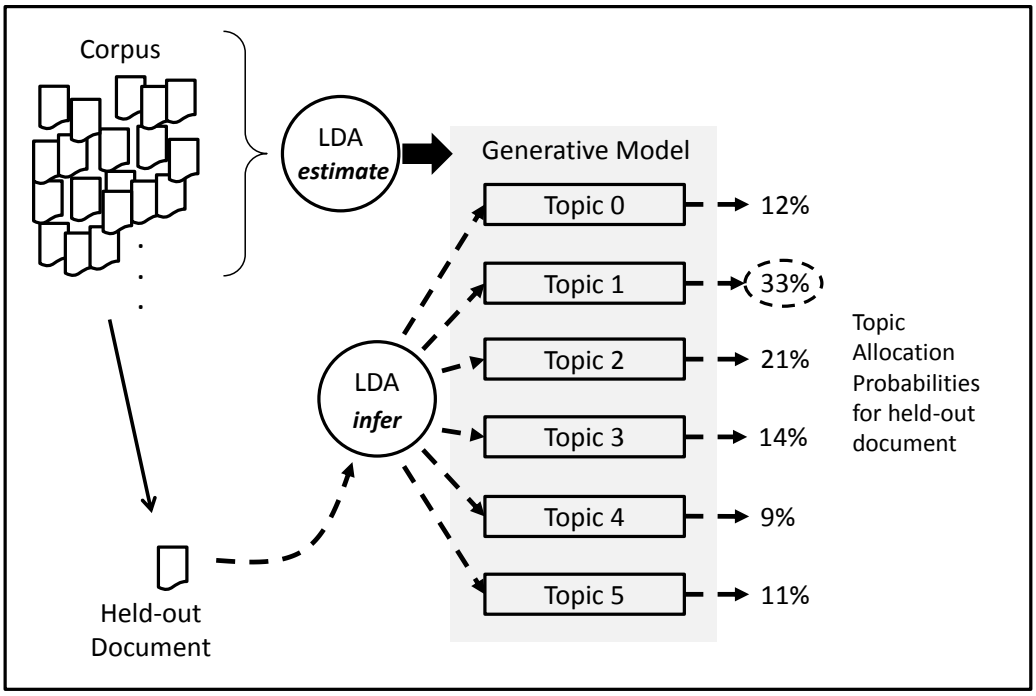

Figure 2. LDA Overview

Overview of the data set

The CEE 181/400/581 instructional team believes in the power of essays to assist students in coming to terms with their beliefs about certain complex problems. Further, we find that writing essays helps students understand and communicate the beliefs of others—an important first step in managing any wicked problem. For this reason we assign a significant writing load each semester (for undergraduates, around eight 400-word essays, and one 4000-word term paper; there are similar requirements for graduate students but with higher required word count). The essay prompts force students to grapple with the world they will face after leaving the university—the world they will have the chance to shape should they decide to engage.

The essays we see can be broken into a few categories—each providing a unique challenge to the LDA classifier (an appendix provides a listing of the prompts for essays used in this research):

1. Some essay prompts result in responses that contain a general discussion of several relevant issues followed by recommendations to pursue a specific approach (e.g., how to manage environmental impact of electronic waste, or, how to manage bio-engineered soldiers in civil society). These essays generally require students to provide an overview of topics (technology, ethics, law, etc.) and then conclude by making recommendations. For this reason the keyword content of each essay is similar, but the recommendation must be supported by the argument. For these, a "bag of words" approach (either LSA or LDA) can be problematic since much of the content is identical and the recommendation might be very brief or consist entirely of common stop words (e.g., "we should proceed with X" v. "we should *not* proceed with X").
2. Other essay prompts result in a discussion of one of a number of specific outcomes (e.g., predict the ascendant culture in 100 years, or, discuss how climate change is a carbon cycle management problem). These essay prompts generally lead to up to a dozen arguably "right" answers depending on the angle an author takes, and depending on the approach, such essays could be wildly different (imagine, for example, the semantic differences between two essays outlining

carbon cycle management; one which suggests we promote tree planting campaigns in social media and one which proposes the rollout of carbon capture and sequestration technologies). This kind of essay also poses challenges to "bag of words" grading approaches due to the broad dispersion of terminology present in high quality and responsive essays.

3. There are also essays which require students to define terminology or distinguish between certain terms and provide examples that support their definitions. For example, students must define "technology" or distinguish between concepts like "science" and "faith." In many cases students resort to dictionary definitions and then append a commentary with examples. In other cases, the examples themselves lead to student-derived definitions. The latter are more thoughtful essays, but the semantic content is quite similar.
4. Some essays require students to comment on the complex interactions of technologies and social systems. For example, writing about how the technology of film can depict an integrated view of technology, culture, and the economy, allows the student to easily grasp the idea that movie producers have an agenda that often extends beyond entertainment into social issues. Also, discussing common cultural discourses like "how do you measure sustainability?" frequently serves to reveal *the student's own* agendas. While each student may be responsive to the prompt, their individual essays are generally only related by the specific examples or domains of discourse chosen by the student.

In addition to the wide range of "right" answers, there are a couple other factors which limit the potential success of the LDA classifier in our particular case. One of these is the relative dearth of data. LDA generally requires large data sets, and to date a maximum of approximately 400 essays has been collected for each prompt (each essay is 400 to 1000 words). On average the number is far less (around 50). This is much fewer than is usually used in LDA research and may be too few to establish a sound model. A second factor is that some of the student essays are of poor quality, so instead of contributing to the model, they tend to detract by including misleading terms. These latter factors have driven the development of the "expert vectors" approach that is discussed later.

Still, despite the high number of degrees of freedom in the legitimate answer space, the LDA experiment has generated some interesting trends that are reported herein.

Method and Approach

Beyond the interest in creating a low impact, low cost automated essay scoring solution, there were two primary goals of this research:

1. First, there was a desire to understand the essay prompts and responses better, and determine if LDA could provide new insight into them, or whether or not LDA could even resolve the many similar responses to the essay prompts. This is essentially a question of experimenting with LDA's already well-attested classification benefits and determining if they could assist with the CEE 181/400/581 corpora.
2. Second, there was a desire to resolve the question of whether or not adequate essay scoring could be automated with a smallish, homegrown software package, or whether a larger (and longer) investment would be required to derive these benefits from LDA (and other tools).

The goal of "understanding"

Once the software was built (see Appendix B and related presentation for details), all essays available from CEE 181 and CEE 400 were imported (CEE 581 essays were not investigated for this effort, but will be reviewed at a later date). Corpora were assembled that contained *all* essays from each class (that is,

responses to all prompts), as well as corpora consisting of essays written in response to *individual* prompts. This is depicted in Figure 3. Corpora used in the analysis are described in Appendix A.

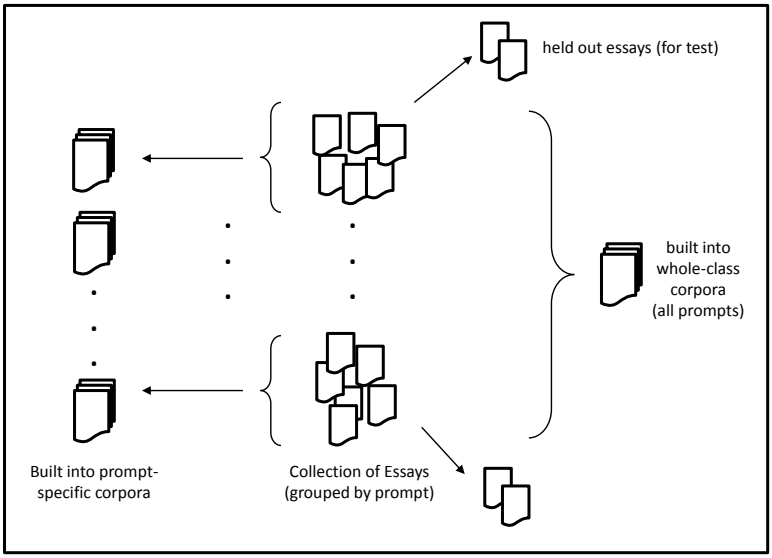

Figure 3. Corpora Construction

Corpora were constructed for whole-class essay groups and prompt-specific essay groups because, as mentioned above, there are many possible answers to each prompt. For this reason, it was important to be able to test the LDA in its ability to not only isolate the within-prompt answer differences (of which there are, on average, around 10) as well as being able to distinguish essays based on the prompt they answered (13 prompts for CEE 181, 9 for CEE 400).

Two outcomes were hypothesized. First, LDA should be able to distinguish and classify the whole-class essay corpora into specific prompt topics (that is, LDA should be able to generate a list of topic keywords that were recognizable by human reviewers to be related to each specific prompt). Second, LDA should be able to distinguish and classify prompt-specific corpora into specific student answer approaches (that is, LDA should be able to generate a list of topic keywords that were recognizable by human reviewers to be valid answers to the specific prompt). Evaluation of the results was expected to provide new insights into the quality of the essay prompts and help the instructional team understand potential overlaps in essay content.

Further, it was of interest to determine if specific instruction regarding the prompts ever resulted in students taking *all* the approaches anticipated by the instructor. For example, the instructions for the "climate change/carbon cycle" essay (see appendix) implies there are at least ten approaches that can be taken in answering the prompt and these general approaches are overviewed for the students to help them see the scope. Seldom, however, are more than three or four of the approaches employed in actual essays.

The goal of "grading"

Wading into the contentious waters of automated essay scoring was facilitated by a very simple hypothesis: LDA topic allocation probabilities (see Figure 2) would (in general) be lower for lower quality essays. That is, bad essays would not very well match the (presumably high quality) essays used to build the model (via the LDA estimation process). This reflects the assumption that lower quality essays are, on the whole, less targeted, more "confused" and therefore more "confusing" to the LDA

classifier. Figure 4 provides a schematic of the idealized hypothetical outcomes. It was anticipated that high confidence allocations would be dominated by high quality essays and the occurrence of "A" grade essays would taper off with a very sharp slope as indicated in Figure 4. Further, it was expected that lower quality essays would be disproportionately represented in the lower confidence allocations. Even in this caricature it is obvious from the significant overlap of the curves (viz. A, B, C, D) that there are complications in differentiating essays by grade. This was mitigated in several ways.

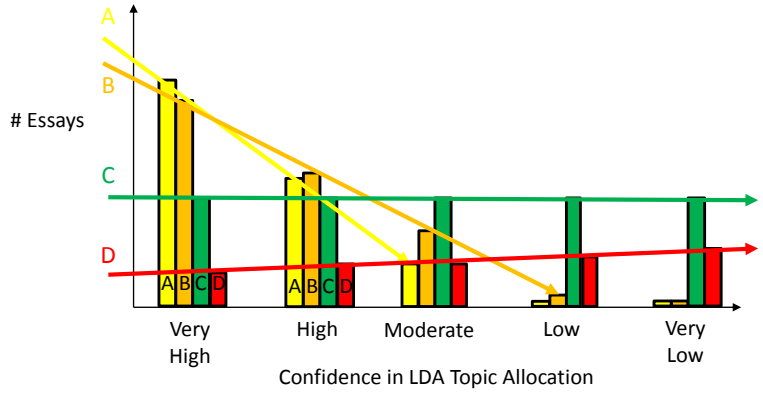

Figure 4. Hypothesis: LDA topic assignment confidence slopes will differ by essay grade

First, models were built on a variety of corpora and inferences were performed on held-out essays (typically 5-10% of the essays). For example, five separate models were built around the "ascendant culture" essay, each holding out a *different* 5% of the essays (see appendix). Piloting was done with the CEE 400 corpus to gain familiarity with the results. Ultimately, the expert vectors were used to evaluate the entire CEE 181 corpus (discussed later). To gain familiarity with the approach and to learn about the quality of the allocations, a human expert rated confidence in the LDA topic allocation for each inference using a five point Likert scale. Table 1 outlines the general approach (0 was reserved for "unassigned").

Table 1. Explanation of Confidence Assignments

Ordinal	Ranking	Explanation
1	Very Low	expected topic is *not* ranked first or second
2	Low	ranked second but greater than ~2% away from first
3	Moderate	ranked second and within ~2% of first
4	High	ranked first but nearest topic is within ~2%
5	Very High	ranked first by greater than ~2%

While occasionally an essay was wondrously confusing, most essays were easily ranked using this approach. Expert judgment was employed when "ties" occurred, but in general, familiarity and long experience with the data set simplified the rankings. It is likely this specific approach would be more successfully employed by someone who was less familiar with the nuances of the essay answers, but since these specific rankings were used to pilot the study and establish trends, and were not directly used in the statistical analysis, this was *not* noted to be a confound.

Second, it was postulated that the standard deviation of *all* the topic allocation percentages (generated by the LDA inference, see Figure 2) might indicate a "spread" or a "confusion" factor that could be used as a quality metric. Figure 5 demonstrates how each essay is allocated to each corpus topic with a certain probability (these sum to 100 percent—not all are shown). The standard deviation of these probability assignments would indicate how "tight" the ranking was. Smaller standard deviations across the

population would indicate a "closeness" or "blurriness" that could be considered a measure of lack of clarity or a famine of appropriate keywords, and hence a lower quality essay, while larger standard deviations would indicate significant "space" between the allocation probabilities. This would imply significant topic isolation and hence better clarity of the essay.

Essay 7930 topics: Confidence: **Very High**

probability	topicName
0.3305490	the anthropocene
0.1876920	sustainable auto products
0.0676920	railroad impact
0.0619780	green chemistry
0.0562640	grid and electric cars
0.0562640	tech clusters
0.0562640	how the world is different from 200 years ago
0.0391210	simple v. complex systems
0.0391210	matrix v. reality
0.0334070	mine operation in developing country
0.0276920	bioengineered soldier

Figure 5. Example of Essay topic allocation probabilities

Figure 6 depicts how this hypothesis might be demonstrated in the data. Note that even with this postulated outcome, it would be difficult to specifically assign a letter grade to specific essays due to the overlap of the curves.

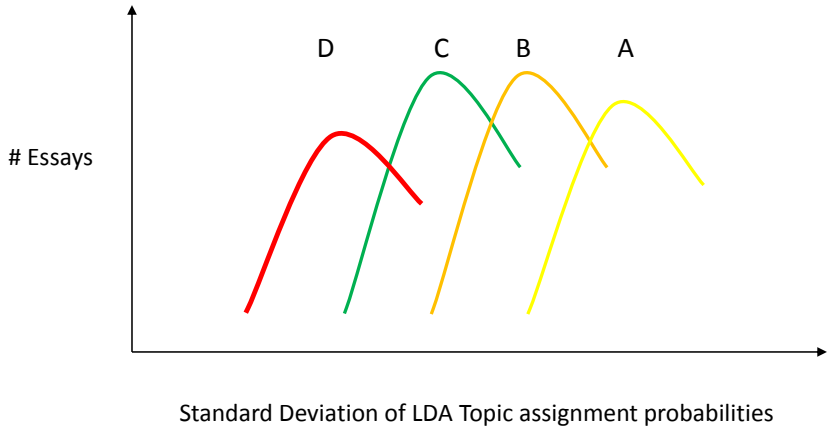

Figure 6. Hypothesis: SD of topic assignment probabilities reflects essay quality

Third, it was noted from the pilot trials with the CEE 400 data set that the *absolute difference* between the *two highest LDA topic allocation probabilities* (i.e., the "top two") seemed more definitive in driving user confidence (again, refer to Figure 5 and note the absolute difference between the first two probabilities in the list is approximately 0.331 – 0.188 = 0.143). For this reason, this measure was investigated as well, and all three of these measurements are pursued in the analysis discussed herein.

Special efforts

Importantly, part of both the insight/understanding and grading goals involved the use of specially constructed "expert vectors." This novelty is discussed here in more detail.

Expert vector corpora were generated from keyword lists and short phrases. Only rarely is a complete expert essay available for our prompts, and even then, such an essay would generally take one of the many allowed approaches to the answer—and specifically *not* address them all. The point of creating the expert vectors was to provide answers that covered *all* the expected angles, but without requiring the significant effort that goes into writing an essay for each. In a sense, the goal was to "game" the LDA. That is, the keyword essays contained concepts that a human grader would be expected to look for in an essay that took a particular tack in answering. For example, the prompt which asks the students to predict the ascendant culture in one hundred years has many legitimate answers, one of which includes predicting a Chinese hegemony for any of a number of reasons. The keyword essay written for this answer is shown here:

> China economy, military
> China economic progress, holds USA debt
> Chinese production and manufacturing
> Growth in population, largest labor force in world
> Manufacturing leadership
> China developing quickly, huge energy producer
> Growing technology producer
> Rare-earth elements
> Chinese people want products, jobs
> Chinese want wealth and economic development
> Chinese spend money in domestic and global markets
> Investing in foreign debt
> Rapid social development

It is obvious that the many factors that lead to a possible ascent of China to cultural dominance are listed here in bullet form without concern for complete sentences or even deep rationale. In fact, the words in this vector could just as well be alphabetized since the LDA classifier does not pay attention to word order. This constitutes an expert vector. Similar vectors were developed for other answers to this prompt, and the many possible answers to all the other prompts. These expert vectors were employed to specifically address the manner in which poorly written essays would dilute the corpus and skew results. In theory, if the LDA is trained on only expert vectors, the ability to adequately grade student essays is greatly enhanced.

Another of the problems facing this research was the limited supply of essays. Because LDA is a statistical approach, larger corpora deliver better results. For this reason, the over-weighting features of the software tool were developed to provide a way to exploit LDA as a high quality classifier even in the absence of huge corpora. The trick is simple: When generating the model, the classifier is directed to include higher quality essays additional times (user-configurable). This allows better quality essays (each a "bag of words") more weight in the model generation process. This is facilitated through the user interface which allows the user to specify the degree of over-weighting that is to be done (see Figure 7). Using this feature, the user can specify that each "A" and "B" essay be included additional times as desired.

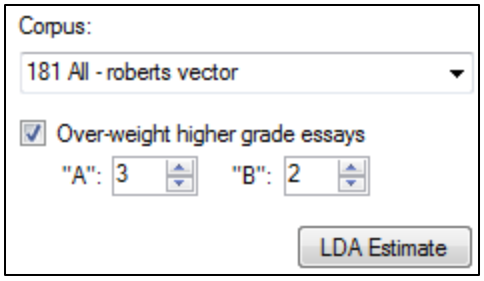

Figure 7. Portion of the user interface that controls over-weighting

As shown in Figure 7, each "A" essay would be included 3 times in a model estimation, while each "B" essay would be included twice. As expected, "C" and "D" essays would be included once each. The success of this feature is still being evaluated and will be reported elsewhere, but the pilot trials on the CEE 400 data set were encouraging. Perhaps obviously, expert vectors would receive "A" grades since these contain the keywords a grader is using to measure quality. Also note that in a corpus generated solely on expert vectors, the over-weighting is not necessary since all the vectors are designed to contain high quality keywords.

Results

As anticipated in the hypothesis, LDA alone is insufficient for grading essays. Complete automated essay scoring solutions will require a variety of machine learning techniques. While LDA is unlikely to be able to do it alone, there are some compelling trends noted in the discussion below. Ultimately, both research goals were satisfied.

The goal of understanding

The first goal (understanding and insight into the data set) was readily accomplished. Table 2 shows that, based on the top word lists, latent topics detected by the LDA reflect the expected outcomes (note, the words are stemmed). Here, for example, it is clear that topic 0 relates to Chinese dominance for reasons including manufacturing and economic growth, while topic 1 predicts continued US dominance for numerous reasons including culture, innovation, etc.

Table 2. Top 15 words for six topics in the "400 Ascendant Culture - roberts vector" corpus

0	1	2	3	4	5
china	usa	natur	india	world	islam
develop	corpor	russia	econom	global	popul
manufactur	econom	lead	market	leadership	growth
debt	leader	leader	workforc	citizen	influenc
progress	innov	gas	reward	unit	domin
strong	cultur	energi	technolog	rise	cultur
quick	busi	signific	respect	govern	muslim
chines	firm	reserv	success	intern	polit
industri	global	power	popul	care	social
lag	lead	largest	progress	human	religion
inform	brand	start	capit	communiti	power
economi	compani	ethic	social	discours	democraci
militari	workforc	hindu	landscap	shift	live
rapid	live	resourc	erad	hegemon	strong
econom	soft	democrat	indian	cooper	birth

The LDA was quite successful at classifying and identifying topics within the data set. Table 3 shows a representative sampling of LDA topic allocation confidence. In general, the LDA inference was "correct"—that is, Very High (VH) or High (H)—93% of the time. In several cases, the LDA pointed out essays that had been incorrectly labeled.

Table 3. Representative Sample of Inference Confidence

Corpus	Inferences	% Very High Confidence	VH	H	M	L	VL
181 All - roberts vector*	703	94%	651	8	8	26	9
181 All - mattick vector	703	90%	604	27	13	29	29
181 All AB only	84	86%	71	1	2	9	1
181 All overweighted	38	95%	36	0	2	0	0
181 All Prompts	29	100%	29	0	0	0	0
400 All Prompts	113	93%	101	4	7	1	0
		Average: 93%					

*Both the "roberts" and the "mattick" vectors had one essay "unassigned" for special circumstances

Interestingly, there were several specific prompts that resulted in significant overlap in essay content. This caused the classifier to "confuse" essays and assign the highest probabilities to the "wrong" topic (here, the use of quotation marks around words like "confuse" and "wrong" reflect human value judgments on an activity the software was doing quite "correctly"). These specific prompts (namely, "impact of tech clusters" and "impact of railroads" in the CEE 181 corpus summarized in the appendix tend to overlap because railroads and their concomitant technologies are an excellent example of a technology cluster) have also been indicted by students for being very similar. When the students confuse the response space, it is likely the software will as well. In this regard, LDA can assist the instructional team in designing better prompts, or at least in understanding some of the student confusion.

The goal of grading

The second goal (analysis of how LDA might facilitate essay assessment) was also accomplished and important trends were found in even the limited data set under study. Both grading hypotheses were validated (see Figures 4 and 6 and compare to Figures 9 and 10).

The analysis presented here was done with the "181 All – roberts vector" corpus which consists of 13 expert vector essays. Note that while other experts would design their answers differently, it was expected that the overlap would be significant and that the results would be similar (this is demonstrated later when the results of the "181 All – mattick vector" corpus are compared). Use of this expert vector allowed all 703 of the student essays to be analyzed without potential confound. That is, none of the student essays were used in the training set for the LDA model estimation (they were all "held out"). This means they all could be used for inference against the model and, hence, be assessed for grading.

Based on analysis of corpus "181 All – roberts vector", Figure 8 demonstrates the trend that confirms LDA's capability to deliver high confidence topic allocations (703 inferences, average absolute difference of top two LDA allocation probabilities=0.23, SD=0.12). Note that nearly all the topic assignments receiving a "very low" confidence are two standard deviations away from the mean. These outliers indicate that the metric of "absolute difference between the top two allocation probabilities" compares well with the user assigned quality of the topic allocation.

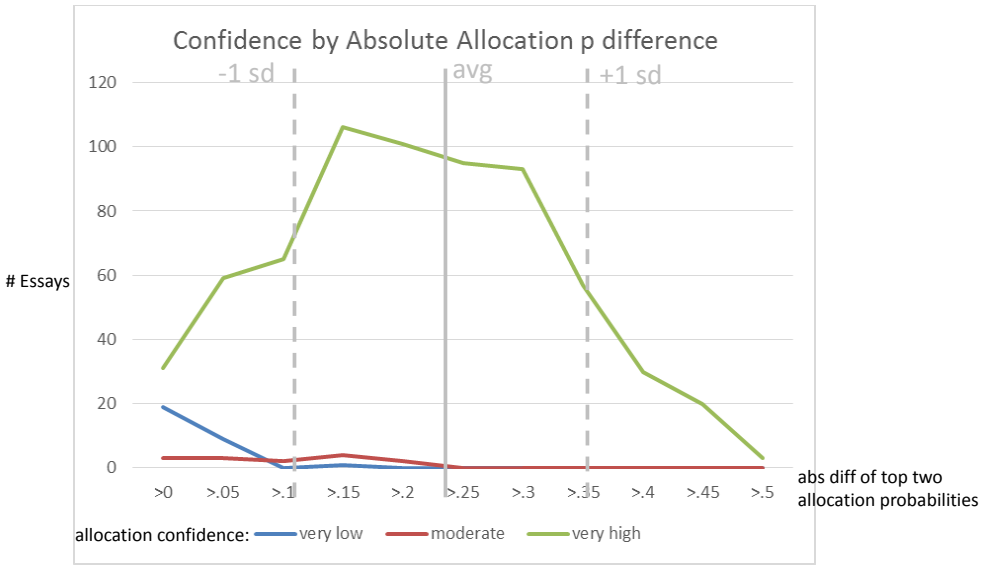

Figure 8. Allocation confidence based on absolute difference of top two topic assignment probabilities

Figure 9 demonstrates the trend that higher quality essays (as graded by human experts) generally deliver higher confidence topic allocations. Note the grades (vertical lines in Figure 9) are overlaid based, once again, on the *average absolute difference of the top two allocation probabilities* (A=0.24, B=0.22, C=0.19, D=0.19). That is, those essays receiving an "A" grade had an *average* absolute difference of the top two allocation probabilities of 0.24 (similarly, this applies to essays receiving B, C, and D grades).

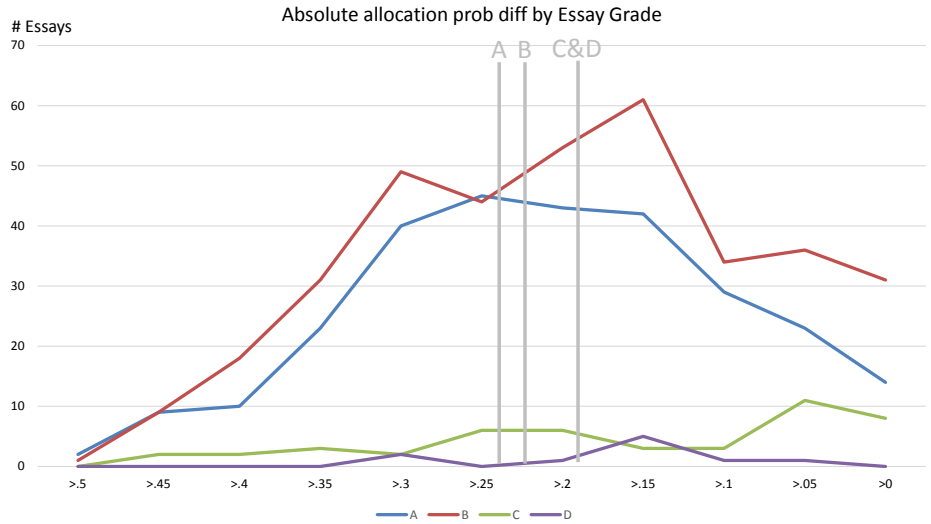

Figure 9. Grades based on absolute difference of the *top two* LDA topic allocation probabilities (vertical lines are positioned at *average absolute difference* for each grade level)

The limited data set prevents real confirmation of the trends predicted in Figure 4, though a fit of the slopes can be observed to reflect the idea proposed. (Note that the data is plotted here with the difference shrinking to 0 on the right, so the slopes would be reversed).

Figure 10 shows the standard deviation of *all* LDA topic allocation probabilities (averages indicated in light gray: A=0.084, B=0.082, C=0.076, D=0.072, SD=0.023). This demonstrates the expected trend that higher quality essays generally deliver higher confidence topic allocations but employs the standard deviation over *all* allocation probabilities instead of focusing on the difference of the top two.

Note that this confirms the second grading hypothesis, though more statistics would be required to derive the isolation of curves projected in Figure 6. Note as well that in Figure 6 the standard deviation is *increasing* toward the right. As plotted in Figure 10 the standard deviation *decreases* toward the right, so the order of curves is reversed. That is, as indicated in the figure, the "A" curve peaks first and the "D" last.

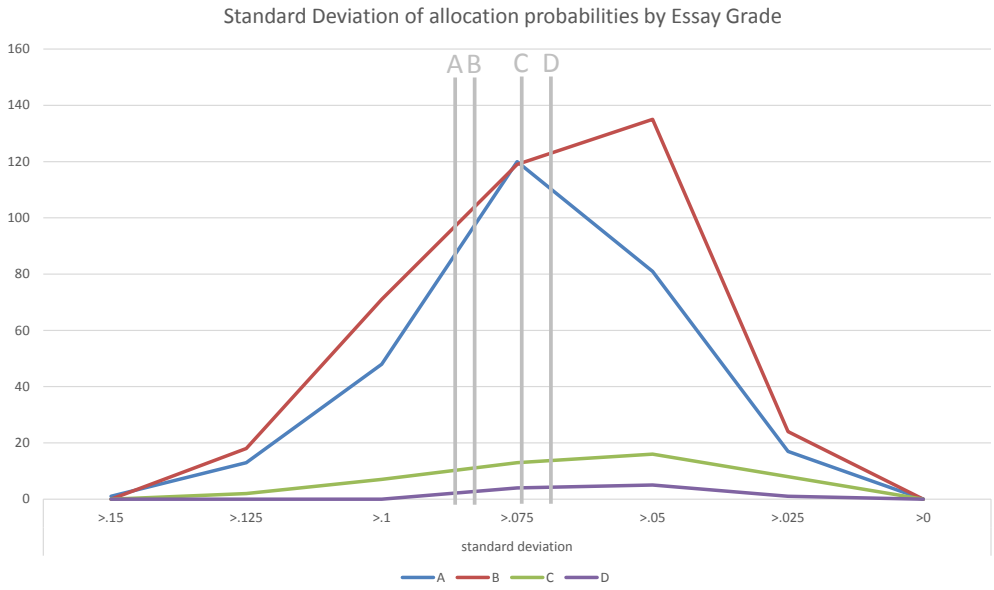

Figure 10. Grades based on standard deviation of *all* LDA topic allocation probabilities (vertical lines are positioned at *average standard deviation* for each grade level)

In the interest of independent confirmation and reproducible results, a similar analysis was done with an expert vector developed by an independent expert (corpus: "181 All – mattick vector"). As shown in Figure 11, this analysis resulted in a nearly identical set of curves when the standard deviations were plotted (averages indicated in light gray: A=0.077, B=0.074, C=0.072, D=0.070, SD=0.026).

It is worth pointing out that the "mattick vector" contained 240 unique words while the "roberts vector" contained 351 unique words. Interestingly, though the "mattick vector" was significantly shorter than the "roberts vector" there was no appreciable difference in the overall outcome of the analysis. This implies that the count of keywords is not as important as the selection of the keywords. This could be a fruitful area for further investigation. Note, however, that the "roberts vector" resulted in a significantly higher number of user-assigned "very high" confidence assignments (651 v. 604, see Table 3), and the "mattick" vector significantly more "very low" confidence assignments (29 v. 9). This might have something to do with the limited keyword set.

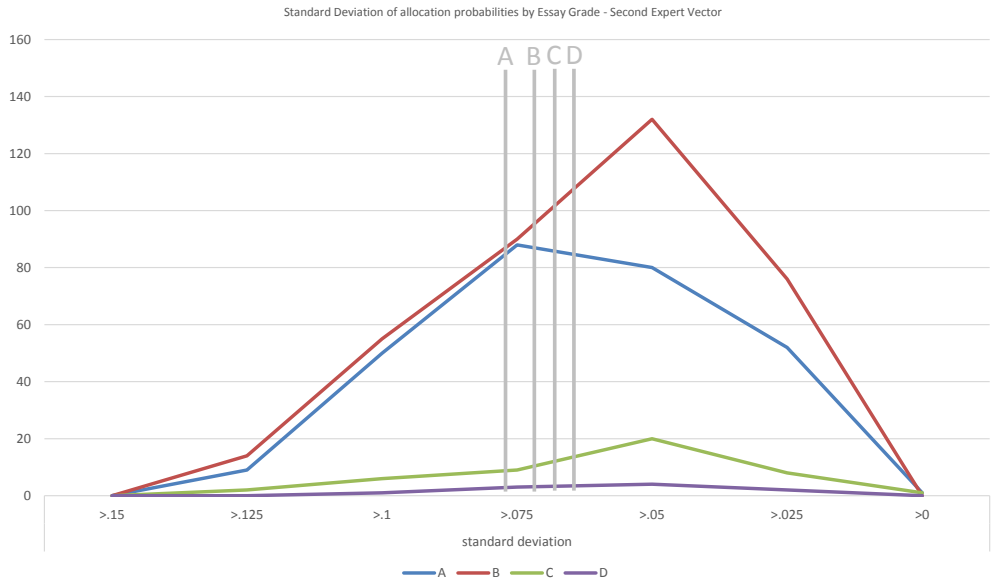

Figure 11. Grades based on standard deviation of all LDA topic allocation probabilities (NOTE: Independent expert vector)

Other details

One confound was identified that involved students who leave the essay prompt in their essay. This tends to augment the essay with a corresponding word set that may have been otherwise under-represented. This can be resolved easily with editing, but was not done for this study due to the effort involved in editing so many essays.

Further, it is clear that including a "references" or "works cited" section artificially inflates the vocabulary of each essay. References contain many important keywords since paper and journal titles are purposefully dense with meaningful words. Further work should be done and comparative studies performed that analyze the influence of these confounds.

Finally, it should be noted that some essays were ranked high by the LDA because they were "on prompt" and used a significant number of required keywords, but were still graded very low because they were unacceptable for other reasons (too short, submitted late, excessively poor grammar, etc.). In the future, these details might be tracked separately to provide higher significance in the data.

Future research might include training a multi-layer neural network on this highly overlapping data and see if it could manage the non-linearity with a Least-Mean-Square approach.

Discussion

It is clear that LDA alone is insufficient for a complete essay grading package. At the very least, however, LDA can point out essays that *largely* answer a prompt in a manner that the instructional team might find "responsive." That is, based on the terminology employed by the student, LDA is quite useful in categorizing an essay as an answer to a particular prompt that reliably reflects the desired keywords (as opposed to an essay that is about some other random topic). This is an important outcome since it means that the LDA toolkit could be employed to validate *minimum acceptable content requirements* for the student essays. If the instructional team finds that the topic allocation probabilities are adequate to classify

an essay as “responsive” in this manner, it might mean that that the “content” requirement can be managed electronically (automatically) and that lower cost essay graders can be employed to worry about the grammar, spelling, and sentence structure elements of good writing.

Conclusion

A custom LDA toolkit was built by incorporating freely available software into a simple user interface that supported the workflow of the instructional team. The software was piloted on the CEE 400 corpus and data analysis was performed on the CEE 181 corpus and reported. The toolset satisfied the two goals of the effort by increasing understanding of the corpus and demonstrating a method by which expert vectors can be used for grading. The research validated the hypothesis under investigation: LDA topic allocation probabilities were, in fact, lower (on average) for lower quality essays. This was demonstrated with several mechanisms including (1) calculating the absolute difference between the top two LDA topic allocation probabilities and (2) the standard deviation of all the LDA topic allocation probabilities. The research did, however, underscore the notion that this trend is not enough to assign a specific grade to a student essay. Still, the trends found in even the small dataset under investigation lend themselves to future investigation.

Acknowledgements

Thanks to Carolyn Mattick and Kurt VanLehn for their helpful feedback on a draft of this paper.

Appendix A: Data Sets

CEE 181 Data Set

Topic Tag	Abbreviated Essay Prompt	Essays
the anthropocene	Why do you think scientists increasingly refer to our modern era as the "Anthropocene"?	59
technology clusters	Why do you think technology clusters have institutional, social and cultural effects, rather than just economic impacts?	53
sustainable cell phone	Explain how you intend to design a "sustainable cell phone."	57
simple v. complex systems	What do you think are the most important differences between complex and simple systems, and why?	56
railroad impacts	As a US villager in the early 1800's exposed to your first railroad, discuss how many of the changes the railroad subsequently caused you think you could have predicted. Are the important changes technological, or are they economic, political, and cultural?	54
sustainable auto products	As vice president of engineering at an automotive firm, design a process that will lead to more sustainable products and write a summary of your proposed process for the company's annual report.	54
mine operation in developing country	As a responsible major mine operator in a developing country, your mine still is causing environmental changes in local ecosystems. Write an op-ed piece for your local newspaper defending your operation against environmental activists who demand that you be shut down.	56
matrix v. reality	How do you know either reality in The Matrix is more "real" than the other? And if even someone in those realities can't tell which is which, what do you think of the morality of Neo's decision to destroy the Matrix?	54
world different from 200 years ago	In what major ways does the world we live in now differ from the world 200 years ago?	58
grid and electric cars	Why should I need to worry about the grid if all I want to do is buy a plug-in vehicle because it's good for the environment?	57
green chemistry	How would you redefine "green chemistry" so that it would be "sustainable chemistry"?	53
ender's game v. iraq war	Compare Ender's Game with the war in Iraq, with its heavy reliance on robotic ground, sea, and air platforms. Do you still think Ender's Game is science fiction? Why or why not?	50
bioengineered soldier	The Army has proposed a technology package that would bioengineer soldiers to be permanently altered to have 200% greater strength, nervous system function, cognitive capability, and skeletal strength. As a senior Pentagon analyst for emerging technologies, write a memo to the Secretary of Defense outlining the potential implications of this set of technologies.	55

CEE 400 Data Set

Topic Tag	Abbreviated Essay Prompt	Essays
ascendant culture in 100 years	Predict the ascendant culture in 100 years.	285
defining technology	Define "technology" taking into consideration the relationship of technology to culture and human interaction with the physical environment.	347
how sci-fi shapes the future	Discuss how science fiction impacts and shapes the development of future technologies. Use *Old Man's War* as an example.	270
film as means to give	Discuss film as a method for presenting an integrated vision of future	250

integrated view of future	technologies in a social, cultural and economic context, using *The Matrix* as an example.	
implications of merging ICT and human consciousness	Discuss the implications of human intelligence becoming integrated with software systems, and thus subject to viral attack. Use *Ghost in the Shell* as an example.	338
science and faith	Discuss the difference between faith and science. Provide an example from environmental science.	341
climate change as C-cycle management	Discuss how climate change can be understood as a carbon cycle design and management challenge	340
short term implications of ICT waste	Write a memorandum to the EPA Administrator discussing the short term environmental implications of information and communication technologies.	174
enhanced ICT in urban environment	Discuss the anticipated effects as enhanced ICT capabilities, including autonomic computing, are introduced into urban systems at all scales.	247

Corpora Used in Testing

- Corpus names are abbreviated tags that allow specific reference. Corpus numbers are simply unique identifiers assigned by the database.
- Similarly named corpora may vary in the number of essays based on the random number of essays held-out
- The number of topics modeled is a research choice and generally reflects a "best-fit" based on several trials. Several corpora were duplicated and modeled based on differing number of topics to determine what might be a best fit.

Corpus (#)	Essays	Held out	Topics	Notes
181 All - roberts vector (56)	13*	703	13	This is a special "expert vector" corpus designed to be compared to ALL CEE 181 essays
181 All - mattick vector (60)	13*	703	13	This is a special "expert vector" corpus designed to be compared to ALL CEE 181 essays
181 All AB only (52)	619	84	13	This corpus was made to enable comparison of all the C and D grade essays since these are more rare and rarely appear in the random hold outs
181 All overweighted (55)	665	38	13	This corpus weights the A and B essays greater than the C and D essays by including the A essays *three* times and the B essays *twice*
181 All Prompts (47)	664	29	13	
181 All Prompts – 2 (48)	649	44	13	These (2-5) are alternatives to "181 All Prompts" which held out a different random selection of essays for test
181 All Prompts – 3 (49)	654		13	
181 All Prompts – 4 (50)	660		13	
181 All Prompts – 5 (51)	654		13	
181 All Prompts – duplicate (59)	664		13	This is an exact duplicate of "181 All Prompts" used to compare the differences in the generated LDA models (model differences were negligible—analysis available upon request)
181 Essay 01 (41)	56	3	5	
181 Essay 01 A (43)	16	43	5	181 Prompt 1 only A grade essays
181 Essay 01 B (42)	36	23	5	181 Prompt 1 only B grade essays
181 Essay 01 overweighted (54)	55	4	5	This corpus weights the A and B essays greater than the C and D essays by including the A essays *three* times and the B essays *twice*

Corpus (#)	Essays	Held out	Topics	Notes
181 Essay 02 (44)	52	6	-	
181 Essay 02 A (45)	17	41	-	
181 Essay 02 B (46)	31	27	-	
181 Essay 10 (38)	50	5	-	
181 Essay 10 A (39)	26	29	-	
181 Essay 10 B (40)	22	33	-	
181 tech cluster only (53)	52	0	2	This special test was configured to determine if LDA could discern two distinct approaches taken by students in defining technology clusters: groups of technologies (e.g., computers and communications), and geographical groupings.
400 All Prompts (30)	2458	375	9	Nine (of about a dozen) distinct prompts were used
400 Ascendant Culture - roberts vector (58)	11**	274	10	Ten topics reflects approximately 10 legitimate approaches that can be taken with this prompt. Using more made it difficult to discern specific topics. Using less "cramped" the answer space.
400 Ascendant Culture 100 years (17)	261	13	10	
400 Ascendant Culture 100 years – 2 (26)	262	12	10	
400 Ascendant Culture 100 years – 3 (27)	264	10	10	
400 Ascendant Culture 100 years – 4 (28)	258	16	10	
400 Ascendant Culture 100 years – 5 (29)	261	13	10	
400 Define technology (20)	336	11	5	
400 Enhanced ICT impacts (23)	235	12	5	
400 Env impacts of ICT (25)	164	10	6	
400 Film and integrated vision of future (21)	239	11	-	
400 GITS and implications of downloaded consciousness (24)	323	15	-	
400 Manage C-Cycle (18)	319		8	
400 Op-ed at 5 years (57)	160		12	This "special" essay asks students to suggest what was missing from their university education. Over time, students have identified a short list of ~20 topics with a solidly recurring top ~10. Due to the free Op-Ed approach, the essay content was such that the LDA was able to recognizably identify only about half.
400 Sci Fi impact on future tech (22)	259		-	
400 Science and Faith (19)	326		-	

* Each of the "181 All…" corpora must specifically exclude the 13 "expert vector essays" in the "181 All - roberts vector" and "181 All – mattick vector" corpora when they are estimated.
** Each of the "400 Ascendant…" corpora must specifically exclude the 11 "expert vector essays" in the "400 Ascendant Culture - roberts vector" corpus when they are estimated.

Appendix B: Software

The software package was written in C# for Windows (.NET 4.0) and deployed as a Windows Forms application. The database was developed for Microsoft SQL Server, but deployed on MS-Access 2013 for ease of use and portability.

The software consists of approximately 1500 lines of C/C++ code which implements the LDA model, and approximately 3500 lines of custom C# that became the user interface and provided the toolset on which this research was based.

Use Cases for the custom software development effort:

1. Manage essay documents
    1.1. Clean essays: remove stop words, special characters, perform stemming, etc.
    1.2. Load essays from file (MS-Word doc or text)
        1.2.1. Assign “class”, “prompt”, “year”, “special tags” (for later filtering)
2. Manage corpora
    2.1. Create corpus
        2.1.1. Filter essays by class, prompt, year
        2.1.2. Select essays from list (allow for “hold outs” for testing)
        2.1.3. Describe corpus for ease of reference
    2.2. Append essay documents to extant corpus
    2.3. Duplicate corpus from extant corpus
    2.4. Compare similar corpora
        2.4.1. Optionally map topic names
    2.5. Delete corpus
3. Perform LDA estimation (build model) and inference (matching)
    3.1. Estimate
        3.1.1. Select corpus
        3.1.2. Run LDA estimate (support “over-weighting” concept)
        3.1.3. Review topic selection
        3.1.4. Name LDA-identified topics (for human use)
    3.2. Infer
        3.2.1. Select held-out essay(s) (that is, essays NOT in corpus)
        3.2.2. Run inference
    3.3. Review results
    3.4. Assess confidence

Software used under the GNU General Public License as published by the Free Software Foundation:

Phan, X-H. (2007). A C/C++ Implementation of Latent Dirichlet Allocation (LDA) using Gibbs Sampling for Parameter Estimation and Inference. Retrieved from: http://gibbslda.sourceforge.net.

Bartocha, K. (2007). C# implementation of the Porter2 stemming algorithm as described at http://snowball.tartarus.org/algorithms/english/stemmer.html.

The stemmer operates as expected by generating “stemmed” words that are not always readable but represent multiple words. For example *economi* can match economy, economics, and economical, while *manufactur* can match manufacture, manufacturing, manufactory, etc.

A more extensive look at the software can be found in the accompanying PowerPoint presentation.

Appendix C: Database

Table	Column	Notes
corpus		Maintains the essay ID lists that identify specific corpus
	essayid	
	corpusid	
corpusDescription		Contains a short description of the corpus
	corpusid	
	description	
corpusTopicAssign		Maps essay words to latent topics found in a corpus
	corpusid	
	essayid	
	wordid	
	topicid	
corpusTopicNames		Maintains the user given topic names
	corpusid	
	topicid	
	topicname	
corpusTopics		Maintains all the LDA estimation (model) data
	corpusid	
	topicid	
	word	
	probability	
essayPrompt		Contains the essay prompts
	prompttext	
	promptid	
	class	
inferTheta		Maintains all the LDA inference probabilities
	corpusid	
	essayid	
	topicid	
	probability	LDA topic allocation probability
	allocationQuality	User-assigned confidence in LDA topic allocation
studentEssay		
	essayid	Each essay receives a unique identifier
	promptid	Each essay is written in response to a specific prompt
	essayYear	The calendar year in which the essay was written
	class	The class for which the essay was written
	grade	The grade assigned to the essay by the instructor
	file	The disk file name that contains the original essay
	tag	Used for specific identification of special essay groups
	essayText	The stemmed essay
	processedLength	Stemmed length (words)
	originalText	The original essay
	originalLength	Original essay length (words)

References

Attali, Y. and Burstein, J. (2006). Automated Essay Scoring with e-rater v.2.0. *Journal of Technology, Learning, and Assessment, 4*(3).

Ben-Simon, A. and Bennett, R.E. (2007). Toward More Substantively Meaningful Automated Essay Scoring. *Journal of Technology, Learning, and Assessment, 6*(1). Retrieved from http://www.jtla.org.

Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*:993-1022.

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S. and Blei, D. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In *Proceedings of the Neural Information Processing Systems Conference 2009*, Vancouver, British Columbia.

Condon, W. (2013) Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing, 18*(2013) 100-108.

Deane, P. (2013). On the relation between Automated Essay Scoring and modern views of the Writing Construct. *Assessing Writing, 18*(2013) 7-24.

Deane, P. and Quinlan, T. (2010).What automated analyses of corpora can tell us about students' writing skills. *Journal of Writing Research, 2*(2) 151-177.

Dzikovska, M.O., Nielsen, R.D. and Brew, C. (2012). Towards Effective Tutorial Feedback for Explanation Questions: A Dataset and Baselines. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 200–210. Association for Computational Linguistics.

ETS (2013a). *Automated Scoring of Writing Quality*. Accessed February 19, 2013 from http://www.ets.org/research/topics/as_nlp/writing_quality/.

ETS (2013b). *Automated Scoring of Written Content*. Accessed February 19, 2013 from http://www.ets.org/research/topics/as_nlp/written_content/.

ETS, Pearson, and the College Board (2010). *Automated Scoring for the Assessment of Common Core Standards*. Accessed February 19, 2013 from www.ets.org/s/commonassessments/pdf/AutomatedScoringAssessCommonCoreStandards.pdf

Flor, M. and Futagi, Y. (2012). On Using Context for Automatic Correction of Non-Word Misspellings in Student Essays. *Proceedings of the 7th Workshop on Innovative Use of Natural Language Processing for Building Educational Applications (BEA)*. pp. 105–115.

Grimes, D. & Warschauer, M. (2010). Utility in a Fallible Tool: A Multi-Site Case Study of Automated Writing Evaluation. *Journal of Technology, Learning, and Assessment, 8*(6). Retrieved from http://www.jtla.org.

Horkay, N., Bennett, R.E., Allen, N. & Kaplan, B. (2005). Online Assessment in Writing. Online Assessment in Mathematics and Writing: Reports from the NAEP Technology-Based Assessment Project NCES Report No. 2005-457. U.S. Department of Education, National Center for Education Statistics.

Kakkonen, T., Myller, N., Sutinen, E., and Timonen, J. (2008). Comparison of Dimension Reduction Methods for Automated Essay Grading. *Educational Technology & Society, 11*(3), 275-288.

Kieft, M., Rijlaarsdam, G. and van den Bergh, H. (2006). Writing as a learning tool: Testing the role of students' writing strategies. *European Journal of Psychology of Education, 21*(1) 17-34.

Klein, P.D. (1999). Reopening inquiry into cognitive processes in writing-to-learn. *Educational Psychology Review, 11*(3) 203-270.

Landauer, T.K., Laham, D. and Foltz, P. (2003). Automatic Essay Assessment. *Assessment in Education, 10*(3) 295-308.

Lavrenko, V. (2009). *A Generative Theory of Relevance*. The Information Retrieval Series, Volume 26. Springer. Retrieved 27 April 2013, from <http://lib.myilibrary.com?ID=187649>

Lim, H. and Kahng, J. (2012). Review of Criterion. *Language Learning & Technology, 16*(2) 38-45 Retrieved from http://llt.msu.edu/issues/june2012/review4.pdf.

Mathews, J. (2004, August 1). Computers weighing in on the elements of essay: Programs critique structure, not ideas. *The Washington Post*, p. A01.

Perelman, L. (2012). Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (AES). In Bazerman, C. et al. (Eds.) *International advances in writing research: Cultures, places, measures*. Fort Collins, Colorado: WAC Clearinghouse/Anderson. Retrieved from http://wac.colostate.edu/books/wrab2011/chapter7.pdf

Patterson, N. (2005). Computerized Writing Assessment: Technology Gone Wrong. *Voices from the Middle, 13*(2) 56-57.

Perez-Marin, D., Pascual-Nieto, I. and Rodriguez, P. (2009). Computer-assisted assessment of free-text answers. *The Knowledge Engineering Review, 24*(4) 353-374.

Porter, M. (1980). An algorithm for suffix stripping. *Program, 14*(3) 130-137.

Ramineni, C. (2013). Validating automated essay scoring for online writing placement. *Assessing Writing, 18*(2013) 40-61.

Ramineni, C. and Williamson, D.M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing, 18*(2013) 25-39.

Vojak, C., Kline, S., Cope, B., McCarthey, S. and Kalantzis, M. (2011). New Spaces and Old Places: An Analysis of Writing Assessment Software. *Computers and Composition, 28*(2011) 97-111.